
Troll, Denise A. "Library Information System II: Progress Report and Technical Plan." The Public-Access Computer Systems Review 1, No. 3 (1990): 4-29.

Note from the Editor:

This article has been condensed from a Carnegie Mellon University Libraries technical report--Library Information System II: Progress Report and Technical Plan, Mercury Technical Report Series, Number 3. To obtain a copy of the full printed report, send a check for $5 to: Mercury Documents Coordinator, Administrative Offices, Carnegie Mellon University Libraries, Frew Street, Pittsburgh, PA 15213.

Abstract

This article describes the work at Carnegie Mellon University in library automation and information retrieval systems. Specific projects include: broadening the range of electronic bibliographic resources by adding databases and expanding the range of stand-alone CD-ROM databases; deepening access to book resources by enhancing catalog records, and adding contents information for scientific and technical proceedings and book reviews to the online catalog; designing a new library information system (LIS II) on a hardware and software platform that demonstrates the feasibility of distributed library systems running on UNIX workstations; and building image databases for the delivery of full-text documents.

The Library Information System II provides for retrieval from several DEC VAX servers using Z39.50 layered on TCP/IP, a search engine from OCLC called Newton, a pilot user interface in OSF X.11 Motif, and an authentication system based on Kerberos and Hesiod developed at MIT. The system is being built to existing and proposed standards, and it is designed to be machine independent. A system which distributes databases over a number of file servers will thus be affordable to a wide range of libraries.

This article address a number of technical and design issues and concludes with an outline of the research and development agenda for the coming year.

1.0 Background

In 1988, Carnegie Mellon proposed building the Library Information System II, a state-of-the-art electronic library

capable of delivering a broad range of bibliographic and textual information to students and scholars. LIS II would be a second-generation system of the highly successful Library Information System currently in place in the University Libraries.

In addition to support from the Pew Memorial Trust, the LIS II project also receives support from the Digital Equipment Corporation, the American Association for Artificial Intelligence, the Online Computer Library Center (OCLC), and Carnegie Mellon University.

1.1 General Goals

The four major goals for LIS II are: (1) expand the breadth and depth of library information available over the campus network, focusing first on expanded coverage of bibliographic information and later on the delivery of the full text of documents; (2) to provide more information about the contents of books by indexing and retrieving the table of contents; (3) to use the capabilities of advanced workstations to improve retrieval, interfaces, and reduce the cost of a large scale retrieval system; and (4) to document and disseminate the results of our work so that if we are successful, our innovations can be diffused within academia. This report discusses progress toward each of these goals.

1.2 General Architecture

Moving information retrieval from a mainframe computer to multiple server machines requires considerable planning and changes in hardware and software. A special computer will be used to build LIS II databases, and special machines will be used as database or retrieval servers. All computers on the campus network or with access to the campus network will have access to LIS II. Workstations and X Windows terminals in the University Libraries and workstations in offices and public computing clusters on campus will run the graphical interface currently being built. Users of other personal computers, like the IBM PC and Apple Macintosh, will run a terminal interface similar to the current LIS I interface.

2.0 Improving Electronic Resources

As the new hardware and software platform is being designed and developed, we are making significant improvements in our electronic resources. We are expanding resources in the existing Library Information System, adding stand-alone databases on CD-ROM, and providing more information about the contents of books we acquire.

To expand the breadth of our electronic collection, we have purchased databases from commercial vendors, and are exploring

the production of databases from local resources. We are also negotiating with publishers to acquire machine-readable journals and technical reports. To expand the depth of our collection, we have designed and implemented several projects to enhance our catalog records for books and technical reports. Each of these developments is discussed briefly below. Whenever possible, additions to the collection are made available to campus as quickly as possible through the current Library Information System, LIS I, so that usage and impact can be monitored and thus contribute to the design of LIS II.

2.1 Expanding the Breadth of the Electronic Collection

We have broadened the scope of our electronic collection by purchasing commercial databases, by acquiring machine-readable text to be mounted locally as databases, and by designing a system architecture that will facilitate the integration of locally produced databases, e.g., Carnegie Mellon administrative databases, into LIS II.

+ Page 7 +

2.1.1 Commercial Databases

To make the best use of our human resources, while developing the distributed retrieval architecture detailed in Section 3, we limited the addition of commercial databases available through the Library Information System to those needed for user tests and planning. We purchased INSPEC (Information Services for Physics, Electronics, and Computing), 1987-present, on magnetic tape and released it to campus (LIS I) in November 1989. INSPEC corresponds to four printed publications: Physics Abstracts, Electrical and Electronic Abstracts, Computer and Control Abstracts, and Update on Information Technology (IT Focus). INSPEC was well received by the physics and engineering communities at Carnegie Mellon. More than 1,700 searches were conducted in this database in May 1990, with an average of 1,900 searches per month since January. Transaction logs of INSPEC searches were used to construct a model of how users search a large, complex database (see Section 3.1.1.4 "Search Complexity and Performance" for details).

In the interest of immediate improvements in resource availability and recognizing that not all databases need to be online on the campus network, we expanded our electronic resources by acquiring a number of CD-ROM products. Eventually we want to provide network access to CD-ROM databases, with the delivery mechanism transparent to the user. The following CD-ROMs have been added to the University Libraries' collection since July 1988.

+ Page 8 +

Table 1. CD-ROM Databases Added Since July 1988

CIRR (May 1990)
Bibliographic citations and abstracts of company and industry research reports provided by securities and investment firms.

Art Index (April 1990)
Bibliographic citations of journal articles, yearbooks, and museum bulletins in all areas of art.

Compact Disclosure (April 1990)
Financial and management information on public companies.

COMPENDEX (April 1990)
Citations of articles, conference papers, and monographs in all aspects of engineering and related areas.

PAIS (April 1990)
Bibliographic citations of journal articles, books, and government documents in public affairs.

COMPUTSTAT (March 1990)
Financial and statistical information on public companies.

CD-MARC (October 1989)
Library of Congress subject authority file and subject headings.

MathSci (October 1989)
Reviews and citations of the world's research literature in mathematics and related areas.

NTIS (September 1989)
Bibliographic citations and abstracts of government-sponsored research and development reports.

+ Page 9 +

The following CD-ROMs are also available.

Table 2. Other CD-ROM Databases

CIS Masterfile (Test Copy)
Bibliographic citations and abstracts of congressional publications.

Statistical Masterfile (Test Copy)
Bibliographic citations and abstracts of statistical information from various publishers.

Social Science Citation Index (Test Copy)
Bibliographic citations of journal articles in the social

sciences.

PsycLit (March 1988)
Journal article citations and abstracts in all areas of psychology.

ABI/Inform (January 1988)
Journal article citations and abstracts on business.

Dissertation Abstracts OnDisc (August 1987)
Bibliographic citations and abstracts of dissertations in all subject areas.

Books In Print Plus (July 1987)
Bibliographic citations of books (in print and forthcoming) in all subject areas.

ERIC (July 1987)
Bibliographic citations and abstracts of journal articles and research reports in education.

2.1.2 Machine-Readable Text

In preparation to begin experiments with the delivery of full-text documents, we are acquiring machine-readable journals and technical reports in the subject field of computer science. We have negotiated with several leading publishers to include their materials online. Elsevier, Pergamon, and the Association of Computing Machinery (ACM) are willing to give us access to their materials. The ACM has committed to providing machine-readable versions of four of its publications: Computing Reviews (10 years), Collected Algorithms (25 years), Communications (2 years), and Guide to Computing Literature (10 years). We have been approached by the Institution of Electrical and Electronics Engineers (IEEE) to provide storage and access to their entire collection of journal page images, over 30 CD-ROMs per year, indexed through INSPEC, and are working on electronic publishing with the American Association for Artificial Intelligence (AAAI). In addition, we are working with MIT, Stanford University, University of Illinois, and the University of California to collect machine-readable computer science technical reports (see Section 3.4 "Developing Standards and Sharing Resources" for details). These materials will be mounted locally as databases.

2.1.3 Local Databases

The success of the Library Information System (LIS I) has stimulated the demand for more online access to campus information. In response to this need, the University Libraries have set the goal of becoming a general electronic publisher for Carnegie Mellon. We intend to provide online full-text databases

of campus information and online ordering of specific services (e.g., ordering textbooks or audio-visual equipment and putting books on reserve) to create an infrastructure for improving support for instruction in the University. As a first step in this direction, we mounted the Faculty/Staff Directory and the C-Book (the student directory) as a database called Who's Who at CMU and released it to campus (LIS I) in February 1989; Who's Who accounts for approximately 8-11% of all searches in LIS, ranging from 5-8,000 searches per month, during the academic year. Plans to mount additional full-text databases are discussed in Section 3.2.2 "Full-Text Databases."

2.2 Expanding the Depth of the Electronic Record

Bibliographic records, originally designed for card catalog use, continue to be the primary access to book collections for users of online catalogs. However, research indicates that the new technology has changed information-seeking behavior, with the result that users are essentially using new search strategies with old information structures. For example, users do more subject searching in online catalogs than they did in card catalogs, and are finding the information in bibliographic records inadequate to their needs--it is often insufficient to retrieve the record or to judge the book's relevance even if the record is retrieved. According to Richard Van Orden, enriching catalog records with information about the content of books may be the next major improvement in information retrieval. Enhanced information can expedite both the remote selection of material and document delivery. The ultimate purpose of catalog enhancements is "the timely provision of selected full-text materials to individuals when and where they need them." [1]

Adding information about the content of books to our online catalog will increase the number of records retrieved and allow users to make better judgments about the value of a book for their particular query. University Libraries have several projects underway to expand the depth of content information available in the online catalog. Some record enhancements have been done entirely in-house and released to campus in LIS I. Two other enhancements have been acquired from commercial vendors and implemented but not yet released to campus: book reviews from Choice, and analytics for books and conference proceedings from ISI (the Institute for Scientific Information).

2.2.1 In-House Catalog Enhancements

Barbara Richards, Alice Bright, and Terry Hurlbert implemented the Online Catalog Enhancements Project in the spring of 1989. The first stage of the project thoroughly examined sample contents pages to determine which kinds of material and how many of each kind should be included in an enhancement project, and to assess the problems that might occur. Based on this review, the

cataloging staff established criteria for enhancing books using definitions of works to be included and works to be excluded; the criteria are discussed below. The review suggested that, provided scientific and technical conference proceedings were excluded, only 25-30% of the new books purchased would qualify for adding table of contents information.

+ Page 12 +

2.2.1.1 Criteria for Enhancement of Catalog Records

o If the contents of a book can be cited separately, then the record is enhanced. Anthologies of plays, collections of critical essays written by different authors, and separately authored chapter titles are three categories of enhanced books. However, proceedings of scientific and technical conferences are excluded from this enhancement for two reasons. First, the length of the tables of contents may exceed a hundred titles, requiring extensive inputting of data, and second, alternative electronic sources, like INSPEC, can provide this information. However, we are placing a flag in conference proceedings catalog records to indicate that the items could be enhanced.

o If the chapter titles within a book provide valuable information about the contents that is not already provided by keywords in the title or subject headings, then the record is enhanced. This category includes chapter titles that delineate historical time periods. Books for which words in the title and supplied subject headings already provide appropriate and sufficient access are excluded. If no unique keywords exist in the contents to improve the description of the monograph beyond the standard cataloging information, then the record is not enhanced; this decision is made by the cataloger.

o If a monograph is an exhibition catalog, then the record is enhanced for each exhibitor whose work is included in the exhibition, with the exception that any exhibition catalog containing more than 25 artists is not enhanced. We are placing a flag in records of exhibition catalogs with more than 25 artists to indicate that the items could be enhanced.

o If a Carnegie Mellon computer science or EDRC (Engineering Design Research Center) technical report has an author-supplied abstract less than one page in length, then the record is enhanced by adding the abstract. If the abstract is longer than one page, then the record is not enhanced.

+ Page 13 +

2.2.1.2 Catalog Enhancement Projects

Three enhancements projects were undertaken in-house. The first project, the only review of existing catalog records, is a special service for the Drama and English departments at Carnegie Mellon, which have a great demand for plays. This project is

adding contents notes (MARC field 505) or added entries (MARC fields 700 and 740) for plays in collections with different authors or the same author. The project was begun by reviewing catalog records for American and English drama; 3,857 catalog records were reviewed and 635 works of collected plays were enhanced. The project is continuing with review of Scandinavian, Italian, Latin, Spanish and French drama.

The second project is adding contents notes (MARC field 505) to the records of newly acquired books with separately authored chapters or chapter titles with valuable keyword information (not provided in the title or subject headings), and to art exhibition catalogs with 25 or fewer artists. To date, 1,187 records have been enhanced. We are flagging records that should be enhanced but are currently not being enhanced, e.g., art exhibition catalogs with more than 25 artists, conference proceedings, and unanalyzed series. Enhancing recently purchased books that meet the criteria for enhancement is an ongoing project.

The third enhancement project is adding abstracts (MARC field 520) to CMU computer science and EDRC (Engineering Design Research Center) technical reports. To date, 1,649 of the total 1,832 technical reports cataloged have been enhanced. The technical reports that were cataloged but not enhanced either had no abstract or the abstract exceeded one printed page in length.

The Online Catalog Enhancements Project has enhanced a total of 3,471 catalog records since October of 1989. Though the project is ongoing, a sufficient number of records have been enhanced and made available online in LIS I to begin studying the effects of these records on retrieval and browsing, i.e., on users' access to information and their ability to discriminate between relevant and irrelevant information. We are collaborating with OCLC to investigate the effects of these catalog enhancements (see Section 4.3 "Research Plans" for a brief overview of our plans).

2.2.1.3 Sharing Enhanced Catalog Records

At the present time, the contents information input by Carnegie Mellon Library staff is only useful to our clientele. The enhanced records created in this project, although created on the OCLC system, are not available to other libraries. Discussions led by Tom Michalak at the February and May 1990 Users Council meetings at OCLC suggest that while many libraries are interested in the potential of enhanced catalog records, support for including records "enhanced" with contents information in the OCLC cataloging system is not yet widespread. However, it seems reasonable that OCLC should allow the contents information input by member libraries to be made available to other libraries who may wish to add such in formation to their catalog records. Unquestionably there will be technical problems which will have to be solved if libraries are to share enhanced records, and Carnegie Mellon will continue to raise the issue of sharing

enhanced records in national databases.

2.2.2 Commercial Catalog Enhancements

Though our in-house record enhancement projects address certain information needs, technical and financial constraints limit what we can do in-house. For example, works with several hundred author-title entries, like conference proceedings, are too costly for an individual library to catalog and the resulting records with contents notes are too large for current systems to handle. One alternative is to purchase analytic records for these items from a commercial vendor and merge these with the Library Catalog. We have two projects of this type underway; the effects of these enhancements will be evaluated along with the in-house record enhancements (see Section 4.3).

+ Page 15 +

2.2.2.1 CHOICE Catalog Enhancements

Choice is a basic book reviewing service for academic and public libraries, emphasizing scholarly titles in their reviews. Choice reviews are available in machine-readable form. Current plans are to make selected records from the Choice database, specifically those that review books in our collection, searchable in our Library Catalog. These records will be searchable along with the catalog records, so that a search for a book title, for example, will retrieve two records--the catalog record for the book and the Choice record with the book review. We modified the Choice records slightly for inclusion in the Catalog. For example, we removed the prices for hardback and paperback purchases, and appended the HOLDINGS field from the Library Catalog record for the book being reviewed to the Choice record reviewing that book. At present, we estimate the addition of 4,000 records to the Catalog using this enhancement for the past three years of the Choice database.

The decision to provide searchable book review records, rather than a hypertext link between the Catalog and Choice records that could be traversed once the bibliographic record was displayed, was a conscious one; its impact on retrieval will have to be measured. We assume that searching book review records with catalog records will facilitate recall of materials, but we do not know if it will facilitate precision and relevance judgments. We will do a cost-benefit analysis after releasing the Choice records to campus. Perhaps later, as an additional test of usage, we will release the entire Choice database as a separate database in LIS II.

2.2.2.2 ISI Catalog Enhancements

Similar to the Choice enhancement project, we plan to include selected ISI (Institute for Scientific Information) analytic and

full records, for books and conference proceedings in science and engineering, in the Library Catalog. Again, we appended the HOLDINGS field from the Library Catalog record for the item indexed in ISI to the ISI record for that item. The analytic records will be searchable, and have a hypertext link to the associated full record with table of contents, which will be displayable from any analytic record. In contrast to the Choice project, where the review record was searchable along with the catalog records, we chose not to make the ISI full table of contents records searchable because all of the information they contain is available in the individual analytic records. We estimate the addition of 15,000 analytic records to the Library Catalog using this enhancement, indexing approximately 1,000 scientific and technical conference proceedings.

+ Page 16 +

3.0 Retrieval System Development

The technical goal of LIS II is to produce an affordable library information system for networked campuses, which are evolving across the nation and the world. Realistically, if libraries are to deliver documents to scholars at their desks, the storage, retrieval, and delivery of information must be cost effective. Furthermore, if libraries are to share electronic resources like enhanced records, we need a communication protocol that supports shared access to information. The goal is to build, not an experimental system, but a hardware and software platform that demonstrates the affordability and usability of the system for campuses of any size. Success depends on establishing standards. The LIS II development team is committed to using established standards, and when development mandates changing or extending standards, to do so within the proper forum for implementing such standards. See Section 3.4 "Developing Standards and Sharing Resources" for details.

LIS II is based on the Andrew system at Carnegie Mellon, developed in a partnership with IBM. Named for both Andrew Carnegie and Andrew Mellon, Andrew encompasses the campus network--in reality a network of more than fifty local area networks, a distributed file system with hundreds of file servers, and thousands of high-function workstations. Workstations facilitate working with multiple applications by providing a window for each application and a window manager to manipulate the application windows, which can be tiled or stacked to produce a two- or three-dimensional workspace, or iconified (shrunk to a graphic) to clear the electronic desktop. These features provide a common user interface to network services, including electronic mail and bulletin boards, printing, and access to the Library Information System. Users can also access the Internet from Andrew, extending their research and collaborative efforts beyond Carnegie Mellon.

+ Page 17 +

The Open Software Foundation (OSF), a non-profit research and development company sponsored by many of the world's major computer firms, recently incorporated the Andrew File System (AFS) into its Distributed Computing Environment (DCE), indicating the acceptance of AFS as a distributed file system standard. OSF distributes a software toolkit and interface style guide that, packaged with the mwm (X.11) window manager, comprise the graphical user interface standard called Motif. Motif has achieved wide acceptance as a standard among hardware and software vendors, and the body of applications implemented with the Motif toolkit, running under mwm, and conforming to the Motif style specifications is growing. Carnegie Mellon has adopted Motif as the campus standard, and the Motif Window Manager (mwm) will be the default window manager for workstations in the Fall 1990. The LIS II development team has adopted Motif as the library standard for user interface design. The result will be a single interface that brings together local applications and services with new third-party software, running across a wide range of machines.

The following two paragraphs provide an overview of our current status and future plans. The rationale and details of each phase of the project are discussed in the sections that follow.

To date, we have created a reasonable model for libraries to share resources under a common interface and demonstrated that the OSI Z39.50 protocol can work across separate servers. The Z39.50 information retrieval protocol allows an application on one computer to query a database on another computer; it specifies procedures and structures for submitting searches, transmitting database records, and access and resource control. An alpha version of basic software components for LIS II was demonstrated at the EDUCOM conference in October 1989. This demonstration included retrieval from several servers across the NSFnet using Z39.50 layered on TCP/IP, a new retrieval system from OCLC called Newton, and a pilot user interface for workstations written in DecWindows. Since then we have added a generalized authentication scheme based on the Kerberos system, converted the user interface to OSF X.11/Motif, and begun name service using Hesiod.

Meanwhile, work has continued on the next phase of the project. By the 1990 EDUCOM conference, we hope to be able to demonstrate storage, retrieval and display of bitmapped images using Fax Group 4 formats. The first work with compound documents, using SGML (Standard Generalized Markup Language) and CDA (Compound Document Architecture), will follow shortly thereafter. During the next year, we will implement LIS II on a new generation of small RISC servers supported by major vendors. This will bring the price of a minimal campus retrieval system to below $100,000, which is considerably less than the cost of running information retrieval on a mainframe. The same technology can be extended to CD-ROM if CD-ROM producers accept networking standards. Though

many vendors are still reluctant to support standards, and licensing restrictions limit networking, we expect to integrate some of our CD-ROM databases into campus networking by the end of 1991. Future work also includes the development of a simple user interface for other personal computers, and a method of statistically monitoring usage.

3.1 User Interface Design

A quality user interface is critical to the success of LIS II. Quality storage, indexing, and retrieval will only enable users to access the breadth and depth of our electronic collection if the user interface supports the tasks they want to do. This phase of LIS II development focuses on building a single workstation interface following OSF's Motif Style Guide. Developing a graphical interface for workstations using the Motif toolkit enables us to overcome some of the problems in interface design encountered with LIS I. For example, users sometimes lost their context when they were working with the VT100 display of LIS I--the only interface available, which responded to each user action by displaying a panel that replaced the panel that prompted the action. Motif offers multiple windows, one for each conceptual task, enabling users to keep their context and build a better conceptual model of information search and retrieval online.

+ Page 19 +

3.1.1 User Studies

In conjunction with implementing a dynamic user interface in Motif, we have analyzed transaction logs, done protocol studies, and conducted lengthy interviews in a wide range of research areas to understand the human factors involved in online information retrieval. The remainder of this section discusses several of these projects, specifically the requirements for journal information, the sequence in which information fields are displayed, the problem of library jargon, and search complexity and performance. Plans for future user studies are included in Section 4, "Research and Development Agenda."

3.1.1.1 Requirements for Journal Information

We have spent considerable time exploring the special requirements of journal and conference information. Journals and conference proceedings often have long titles that, when truncated for the one-line-per-record display, become meaningless, e.g., "International Journal of." Furthermore, journal titles often change over time as the journal is re-named to better identify its contents in a changing discipline or to reflect a merger with another publication; these name changes create considerable problems for users and are difficult to track in systems without cross referencing and linked records.

Indications of journal holdings are also problematic because subscriptions are sometimes intermittent, issues are sometimes missing, and information about the most recent issue is often not entered into the system in a timely way. Additionally, since our journals are for the most part shelved alphabetically by main entry (which is not necessarily the same as the title) rather than by assigned call number, users often have trouble locating the journal even when they know we have it in our collection. To complicate matters still further, LIS I transaction logs indicate that users clearly want to search the contents of journals for author, title, and subject information, not just search a database of journal records to see if we have a journal.

The results of our research on journals to date indicate that LIS II should provide the following:

o a one-line-per-record display that includes meaningful (usable) information

o a brief record display that includes variant journal titles and Carnegie Mellon holdings

o a full record display

o an item- or issue-level display that includes real-time updates of latest issues

o a table of contents display accessible from the issue-level display

o a simple way to track journal title changes

o a display for browsing variations of journal titles

o links between records in other databases (e.g., INSPEC) and associated journal records

o a simple way to request a photocopy or FAX or to submit an interlibrary loan request

3.1.1.2 Sequence of Displayed Fields

Since many database records are several screens long (in LIS I) and research indicates that users often do not display more than the first screen, the sequence in which information fields are displayed is very important to user satisfaction. Traditionally, our catalog records have displayed information in a sequence suitable for librarians or system designers, but not necessarily suitable for patrons of the electronic library. For example, esoteric information fields like 008, CODES, ACQNUM, and DOCNUM are displayed at the top of the record, while the information fields that users typically use are displayed farther down the

record, often interspersed with more esoteric or less important fields, e.g., LC-CARD and LANGUAGE (usually English). This sequence results in users having to scan the full records for relevant information, which may be displayed on subsequent screens. Our goal is to reorganize the sequence of fields so that those typically used by library patrons are at the top of the record and thus appear on the first screen when the full record is displayed.

3.1.1.3 Library Jargon

Another study examined jargon in library handouts and reference interviews (in preparation for online searching). The results of the study reveal that patrons misunderstand library terms approximately half of the time. The implications for LIS II are far reaching, not only in terms of the language to be used in the online help and on the buttons and menus, but in terms of what tags or labels to attach to the different information fields in the records themselves. For example, in a multiple choice test, only 35 out of 100 test subjects (CMU freshmen) selected the correct definition for the term "citation"; most subjects drew on their knowledge of parking or speeding violations and defined "citation" in the library context as a notice of overdue books. At present, "citation" is a tag we use to identify a field in our Library Catalog records; obviously this tag does not communicate effectively to everyone.

3.1.1.4 Search Complexity and Performance

Using transaction logs for INSPEC, we created a model of user searches to use as a base line for preparing LIS II. We examined the logs from one of the busiest afternoons of the academic year to determine the following:

o the number of searches issued per minute

o the number of users on the system simultaneously

o the complexity of user searches, defined as a function of the number of terms per search; the use of Boolean and proximity operators, field restrictors and truncation; and instances of browsing or scanning the index

LIS II will handle 25 simultaneous users generating searches at the same rate and complexity found in LIS I. The goal is to provide performance that exceeds current LIS I performance on 70% of real searches. Users entering searches that exceed performance guidelines by 50% will be given a resource control option to cancel or proceed; if the search exceeds the guidelines by 100%, the user will be given an option to cancel or browse the index to narrow the search. Resource control is discussed in Section 3.3 "Distributed Retrieval Architecture."

3.2 Database and Document Types

One of the goals of LIS II is the delivery of complex documents over the network. While the current implementation of LIS II supports only ASCII text, both the full text of documents and structured information such as bibliographic records, over the next few years, the formats and sources of data available in LIS II will increase. The focus of our research in this area is on image databases, full-text databases, and personal databases.

3.2.1 Image Databases

Our first priority is to extend the architecture of our entire computing environment so that it supports bitmapped images as well as ASCII text. Information from paper sources, such as journal articles, will be made available in bitmap format (see Section 2.1.2 "Machine-Readable Text"). We will use CCITT Fax Group 4 format to store the compressed images and will provide software decompression and display tools on the individual workstation. This area calls for a wide range of research on storing and displaying images with high resolution, gray scales, and color. Reasonable display performance of bitmaps depends on the speed of the decompression algorithm, the caching of data, and the ability of the decompression algorithm to work ahead of the user interface. The retrieval of bitmap data has implications for the retrieval protocol and requires changes in Z39.50. For example, the application level flow control in Z39.50 is record oriented, but the size of records containing bitmapped images may exceed 50 KB, making it necessary either to retrieve partial records or to retrieve bitmapped images from a secondary server. The format of the data likewise requires special handling by the user interface.

3.2.2 Full-Text Databases

In the future, full-text databases with very different indexing schemes from bibliographic databases will be added to LIS II. As electronic publishers for Carnegie Mellon, the University Libraries intend to provide online full-text databases of the following campus information:

- software licensing and availability information
- career resources information
- Carnegie Mellon policies and procedures manual
- the undergraduate catalog

o Macintosh and Andrew system user help files

o faculty and staff publications and research profiles

o indexes to student and faculty newspapers--perhaps with full text

Additional full-text databases will include research materials as well as standard office reference materials, such as phone books, encyclopedias and dictionaries. Because unpublished working papers and postings on bulletin boards are of vital importance in some disciplines, e.g., computer science, LIS II will merge published and unpublished information. We will provide indexed access to Carnegie Mellon working papers, and make use of work that is being carried out on automatic indexing of Arpanet bulletin boards, so that selected bulletin board postings can also be added to the retrieval servers.

3.2.3 Personal Databases

The original conception of a library information system was to bring a search index to a user as a single isolated tool. Our investigations and interviews with users led to a new conception based on the knowledge that documents are rarely used alone. The new understanding is that retrieval technology is an adjunct to desktop management, therefore a library information system must be integrated into the larger work environment. There is a growing tendency among users to want to leave the library connection active all day rather than log in and out of the application repeatedly; this trend will have a significant impact on established system designs, which commit actual hardware to each connection. With this in mind, we intend to use emerging standards to link LIS II documents to word processors, databases, electronic mail, and similar applications. We will provide toolkits for individual users to make databases available through LIS II. Using the toolkits, personal database creators will be able to access their databases through LIS II, or provide their colleagues with access to their databases through LIS II.

The next challenge in handling document types is storing the source of a document, e.g., author-contributed text in machine-readable form. We are acquiring source documents for future research and development. We will use SGML and CDA to describe the intellectual structure and content of the document and to guide the format of the display. An example of the problems to be solved in this area is the relationship between spreadsheets and tables for display and page layout. A major area for future research, but beyond the current plans for LIS II, is the handling of dynamic documents. Postscript is another format for non-revisable documents, and we are planning support for it.

3.3 Distributed Retrieval Architecture

The distributed architecture of LIS II requires a range of support services. The first is a mechanism to identify and describe databases on the network. Our total Database Information Service requires a number of features in addition to those traditionally provided. The long term goal of the Database Information Service is for users to be able to find information without knowing which database to search. In conjunction with this service, the system requires authentication, access control, and resource control. We need a password and security (authentication) system for multiple reasons. The primary reason is to control access to licensed databases, but we must also limit access to sensitive data within databases, for example, to social security numbers in the Who's Who at CMU database. Additionally, authentication of individual users enables us to collect meaningful statistics about the behavior of different classes of users, e.g., in different disciplines. The Kerberos authentication scheme is used as the basis for this service. Resource control is the final major service required by LIS II. A distributed architecture designed to be used across institutions must include a mechanism for limiting the amount of resources that can be consumed by a remote user. This protects against abuse, makes it possible to provide subscription services for licensed databases, and protects users from potentially costly mistakes by notifying them of expensive requests.

3.4 Developing Standards and Sharing Resources

As an affordable platform for sharing library information, we expect that LIS II will be expanded in the future. To this end, we are working with other groups to develop standards that all libraries can use. This section briefly discusses several projects in this area. See also the earlier discussion of sharing enhanced catalog records (see Section 2.2.1.3).

Members of the LIS II development team participate in the Z39.50 Implementors Group. We are lobbying for extensions to the protocol based on our work with LIS II, where we found it necessary to extend the protocol by devising local conventions:

- for representing Boolean queries
- for using Z39.50 element set names to provide alternate views of retrieved records
- for sorting retrieved records on both the retrieval server and the user's workstation
- for browsing indexes

Further extensions to the protocol may also be necessary, e.g.,

to handle retrieving image data.

Two other projects for testing shared resources are in the planning stages. The first project, with MIT, Stanford University, the University of Illinois, and the University of California, is to build a distributed collection of computer science technical reports and working papers; the result will be a full-text database with the items held at separate locations but with a shared index. Searchable bibliographic records with abstracts will be provided at each site, with the full text stored as page images in an image database at the home site. The second project, with the University of California and Pennsylvania State University, will test extensions of the Z39.50 protocol by sharing library catalog records; this project is sponsored by Digital Equipment Corporation.

Additionally, we are working with Andrew system administrators to implement standards for Motif applications and window management at Carnegie Mellon. This involves collaboration on user testing and document preparation so that interactions and terminology are identical across applications.

4.0 Research and Development Agenda

In conclusion, our LIS II plans for the next year include work in development, implementation, and research. Each of these is discussed briefly below, with the items in each section listed in order of priority.

4.1 Development Plans

o Test the graphical user interface--the number, placement and design of the windows; the text of error messages, buttons, menus, online help; the interactions between searching and browsing; the number and type of indexes to provide for each database; the information to include in the one-line-per-record displays; and the sequence of displayed fields in database records. Several research methods will be used, including protocol analysis, structured interviews, and user questionnaires. The results of these studies will affect the design of the user interface.

o Build a terminal interface for personal computers like the IBM PC and Apple Macintosh. Because of the popularity of the Macintosh at Carnegie Mellon, long-term plans include building a Macintosh interface to LIS II.

o Instrument the system to monitor user behavior based on a profile of significant characteristics--like college, department and status (e.g., Fine Arts, Drama , undergraduate); location where search was issued (e.g., office, public cluster, or library); database selection; search terms (including operators

and restrictors); browse terms; instances of opening and closing windows; the number of short (one line per record) and full records viewed, and the number and sequence of page images viewed, etc.

o Handle complex documents--using SGML and CDA to describe their form and content.

4.2 Implementation Plans

o Implement LIS II on distributed file servers and release to campus.

o Provide training and documentation for library staff and patrons--to facilitate the shift from a terminal emulation interface (LIS I) to a workstation interface (LIS II).

o Broaden the range of bibliographic databases available in LIS II.

o Provide full-text databases--both searchable ASCII text of campus information and reference works, as discussed in Section 3.2.2 "Full-Text Databases," and displayable page images, as discussed in Section 3.2.1 "Image Databases." We are focusing on image databases and will continue experiments with different scanning, scaling, and compression-decompression algorithms.

4.3 Research Plans

o Evaluate the effects of catalog enhancements on recall and precision--preliminary results from a pilot study of the current system (LIS I), planned for Fall 1990, will be used to design a more rigorous evaluation of catalog enhancements in the new system (LIS II). We want to assess the number of additional access points made available in the enhancements, the effects on retrieval, the effects on relevance judgments, the impact on the size of the catalog, and the cost per enhancement. The results of this evaluation should facilitate sharing enhanced catalog records.

o Evaluate and document the transition from LIS I to LIS II.

o Evaluate user behavior and preferences with LIS II--how skills develop over time; how acceptance is influenced by user characteristics, such as social group (student, faculty, staff, alumni) and discipline (engineering vs. social sciences), and by various features of the system itself, e.g., multiple windows, databases, indexes. Results from studies of user characteristics and skill levels will contribute to the ongoing design of the system.

o Study how users use full-text databases--for example, given page images of journal articles or technical reports, do users read the pages sequentially or skip around in the text? This study will entail instrumenting the system to monitor user behavior and running user protocols to better understand why users do what they do. The results of the study will help us develop suitable navigational tools and caching procedures for full-text databases.

+ Page 29 +

References

Van Orden, Richard. "Content Enriched Access to Electronic Information: Summaries of Selected Research," Library Hi Tech 8, No. 3 (1990): 28.

About the Author

Denise Troll
Carnegie Mellon University Libraries
Frew Street
Pittsburgh, PA 15213.
BITNET: troll+@andrew.cmu.edu
